PARNT: A Statistic based Approach to Extract Non-Taxonomic Relationships of Ontologies from Text

Ivo Serra, Rosario Girardi
Federal University of Maranhão, Computer Science Department
São Luís, Maranhão, Brazil
ivocserra@gmail.com, rosariogirardi@gmail.com

Paulo Novais
University of Minho, CCTC/Department of Informatics,
Braga, Portugal
pjon@di.uminho.pt

***Abstract*—Learning Non-Taxonomic Relationships is a sub-field of Ontology learning that aims at automating the extraction of these relationships from text. This article proposes PARNT, a novel approach that supports ontology engineers in extracting these elements from corpora of plain English. PARNT is parametrized, extensible and uses original solutions that help to achieve better results when compared to other techniques for extracting non-taxonomic relationships from ontology concepts and English text. To evaluate the PARNT effectiveness, a comparative experiment with another state of the art technique was conducted.**

Keywords- Learning non-taxonomic relationships; Ontology; Ontology learning; Natural language processing; Machine learning

I. INTRODUCTION

Manual construction of ontologies by domain experts and knowledge engineers is a costly task, thus automatic and/or semi-automatic approaches for their development are needed. Ontology Learning [16] [17] [19] aims at identifying the constituent elements of an ontology, such as non-taxonomic relationships, from textual information sources.

Some techniques have already been proposed for Learning Non-Taxonomic Relationships of Ontologies (LNTRO) [1] [2] [5] [12] [21] [25]. All of them use Natural Language Processing (NLP) techniques [10] [18] to annotate the corpus with the information needed for subsequent processing, as well as Machine Learning (ML) [4] [24] to refine the relationships that result from the previous phases. However, most of the techniques that use ML [2] [12], have limited application since they require corpora with specific characteristics, namely having a great amount of verb phrases composed of only one word or texts structured in titles and text body, as it will be explained in section III.

This article presents PARNT, a novel LNTRO technique that is widely applicable and may help ontology developers to gain more efficiency in this task, because of its higher level of parametrization and more adequate solutions to the refinement phase of LNTRO (section V), which performs a better separation between valid and invalid relationships.

The paper is organized in seven sections. Section II introduces a formal definition of ontology. Section III defines the general process for LNTRO. Then, section IV describes some representative techniques of the state of the art in LNTRO and which solutions they adopt for each of the phases of the general process. PARNT, our proposal for LNTRO, is detailed in section V. Section VI presents and discusses an evaluation of PARNT against another state of the art technique. Finally, section VII presents the conclusions concerning the benefits and some future work on our proposal for LNTRO.

II. A FORMAL DEFINITION OF ONTOLOGY

An ontology is a formal and explicit specification of a shared conceptualization of a domain of interest [26]. Conceptualization refers to an abstract model of some phenomenon in the world. Explicit, means that the type of concepts used and the limitations of their use are explicitly defined. Formal, refers to the fact that the ontology should be machine readable. Shared, reflects the notion that an ontology captures consensual knowledge, that is, it's not private to some individual but accepted by a group. Currently, ontologies are applied in areas such as the communication of software agents [11], integration of information [23], composition of Web Services [5], description of contents to facilitate their recovery [15] in NLP [20], in the Semantic Web [13], in building knowledge-based systems [17] and in applications of knowledge management [8]. Formally, an ontology can be represented by a 6-tuple [19]:

$$O = (C, H, I, R, P, A) \quad (1)$$

where,

$C = C_C \cup C_I$ is the set of entities of the ontology. They are designated by one or more terms in natural language. The set C_C consists of classes, i.e., concepts that represent entities that describe a set of objects (for example, "Person" $\in C_C$) while the set C_I is constituted by instances, (for example "Anne Smith" $\in C_I$);

$H = \{kind\_of\ (c_1, c_2) \mid c_1 \in C_C, c_2 \in C_C\}$ is the set of taxonomic relationships between concepts, which define a concept hierarchy and are denoted by "kind_of(c_1, c_2)", meaning that c_1 is a subclass of c_2. For instance, "kind_of(Costumer, Person)";

$I = \{is\_a\ (c_1, c_2) \mid c_1 \in C_I \wedge c_2 \in C_C\}$ is the set of relations between classes and its instances. For example, "is_a(Erick, Lawyer)";

978-0-7695-4967-5/13 $26.00 © 2013 IEEE
DOI 10.1109/ITNG.2013.70

R = {rel_k (c_1, c_2, ..., c_n) | ∀i, c_i ∈ C} is the set of ontology relationships that are neither "kind_of" nor "is_a". For example "represent(Lawyer, Costumer)" and "represent(Erick, Anne Smith)";

P = {$prop_C$ (c_k, datatype) | c_k ∈ C_C} ∪ {$prop_I$ (c_k, value) | c_k ∈ C_I} is the set of properties of ontology classes. "$prop_C$" defines the datatype of a property while $prop_I$ defines its value. For instance, "subject(Case, String)" is a $prop_C$ element while "subject($Case_{12}$, adoption)" is a $prop_I$ element.

A = {$condition_x$ ⇒ $conclusion_y$ (c_1, c_2,..., c_n) | ∀j, c_j ∈ C_C} is a set of axioms, rules that allow checking the consistency of an ontology and infer new knowledge through some inference mechanism. The term $condition_x$ is given by $condition_x$ = {($cond_1$, $cond_2$, …, $cond_n$) | ∀z, $cond_z$ ∪ H ∪ I ∪ R}. For example, "apply(defense_$argument_{22}$, $Case_{12}$) ∧ similar($Case_{12}$, $Case_{13}$) ⇒ apply(defense_$argument_{22}$, $Case_{13}$)" is a rule indicating that these two legal cases are similar thus the same defense argument can be used in both cases.

III. The General Process for LNTRO

LNTRO looks for automating or semi-automating the extraction of non-taxonomic relationships from text and may be accomplished through the activities described in the process of Fig. 1 [9].

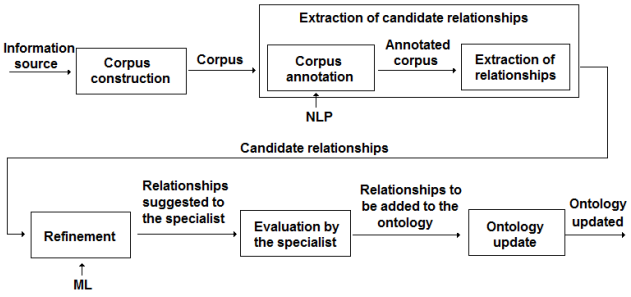

Figure 1. A generic process for LNTRO.

The *corpus construction* task consists in selecting documents of the domain from which relationships can be extracted. This is usually a costly task and the outcome of any LNTRO technique depends on the quality of the corpus used.

The *extraction of candidate relationships* task aims at identifying a set of possible relationships. It has the corpus built in the previous phase as input and candidate relationships as its product. It is composed of two sub-activities: *corpus annotation* and *extraction of relationships*. The *corpus annotation* task tag the text using NLP techniques that are necessary for the next steps of LNTRO. In the *extraction of relationships* activity, the annotated corpus is searched for evidence suggesting the existence of relationships. For example, Maedche and Staab [2] consider the existence of two instances of ontology concepts in a sentence as evidence that they are non-taxonomically related. For Villaverde et al. [12], a relationship is identified by the presence of two ontology concepts in the same sentence with a verb between them.

The relationships obtained from the previous task should not be recommended to the specialist since there is usually a substantial amount of them that do not correspond to good suggestions. For this reason, in the *refinement* phase, ML techniques may be used to try to deliver the best suggestion to the specialist.

In the *evaluation by the specialist* task, he/she selects and possibly edits the relationships to be added to the ontology from those outputted from the previous phase. Finally, in the *ontology update* activity, the file with the ontology is updated with the relationships that were chosen by the specialist.

IV. Related Work

Several LNTRO techniques have already been proposed [1] [2] [5] [12] [21] [25]. In this section some of them are briefly presented and discussed, highlighting the reasons that motivated the development of PARNT.

Villaverde et al. [12] proposed a LNTRO technique that uses part of speech tagging in the *corpus annotation* phase in order to identify the verbs. In the extraction of relationships a tuple in the form <c_1, v, c_2> (c_1 and c_2 are ontology concepts and v is a verb) is generated for each occurrence of two consecutive concepts with a verb between them. For example in the sentence "The young couple have two children." the generated tuple <*couple, have, child*> is a candidate relationship. To possibly increase recall, concept synonyms [3] are included in the search for concepts in the text. In *refinement* the candidate relationships are subjected to an algorithm for mining association rules [22] that represent relationships between concepts. For example, the rule *(couple ∧ child)* → *have* means that there is a non-taxonomic relationship between the two concepts "couple" and "child" and that the label is the verb "have". Support and confidence [22] are two thresholds to prune these relationships and their values are defined experimentally by the specialist.

Maedche and Staab [2] proposed a technique that extracts non-taxonomic relationships from corpora with instances. It uses named entity recognition (NER), in the *corpus annotation* phase, to associate instances found in the text to its corresponding classes. In the *extraction of relationships* phase a tuple in the form <c_1, c_2> (c_1 and c_2 are ontology concepts) is generated for both of these situations: for every couple of instances of ontology concepts in the same sentence and for every instance in a title with every instance in its text body. In *refinement*, the candidate relationships are submitted to the algorithm for mining generic association rules [2] in order to present relationships in the form *(c_1 → c_2)*, meaning that there is a non-taxonomic relationship between c_1 and c_2. This algorithm also suggests the best level in the ontology taxonomy to insert the relationship. For instance, in the case of a supermarket, the algorithm could suggest that "snacks are purchased together with drinks" rather than "chips are purchased with beer" and "peanuts are purchased with soda".

Sanchez and Moreno [5] developed a LNTRO technique that performs the *corpus construction* automatically via queries in a web search engine. Part of speech tagging is

used in the *corpus annotation* phase to identify verb phrases in each sentence. In the *extraction of relationships*, for each sentence, relationships represented by tuples in the form $<np_1, vp, np_2>$ are extracted for every verb phrase (*vp*) with the first noun phrase to its left (np_1) and the first noun phrase to its right (np_2). In the *refinement* phase a statistical solution is adopted, consisting in the execution of predefined formulas to check the degree of relatedness of the extracted tuples *(*$<np_1, vp, np_2>$*)* with the domain.

A technique proposed by Fader et al. [1] uses verb phrase chunking, in the corpus annotation phase, to find verb phrases in each sentence. In the *extraction of relationships*, the relationships represented by tuples in the form $<np_1, vp, np_2>$, where *vp* is a verb phrase and np_1 and np_2 are the first noun phrases to its left and to its right, are generated for each sentence. In the *refinement* phase a logistic regression classifier is used to rank the extracted tuples according to their probability of being valid relationships. Two examples of the characteristic variables used by the logistic function are: the sentence has less than ten words and the noun and verb phrases correspond to the complete sentence.

About the mentioned approaches, Maedche and Staab [2] apply the algorithm for mining generic association rules [2]. Athough it performs the proposed functionality of suggesting the adequate hierarchical level for the non-taxonomic relationships, it is hardly applicable in practice. Villaverde et al. [12] use, in the *refinement* phase, an algorithm for the extraction of association rules [22] that recommends non-taxonomic relationships including its corresponding verbal labels. However, in practice this solution leads to a decrease in effectiveness when compared to algorithms that suggest a list of labels for each pair of concepts like bag of labels (section IV) or even to those that do not suggest any label, like the frequency of tuples composed of two concepts ($<c_1, c_2>$). For these reasons, these two proposals can present good results only in specific situations. It is the case of the experiment conducted by Villaverde et al. [12] that presented reasonable results in extracting non-taxonomic relationships from Genia [7] only because a great amount of verbs in this corpus belonged to a relatively small set of unigram verbs.

Sanchez and Moreno [5] automate the costly task of corpus construction using queries in a web search engine whereas Fader et al. [1] use a logistic regression classifier to rank non-taxonomic relationships according to their probabilities of being valid ones. However, because they do not have ontology concepts as inputs, like Villaverde et al. [12] and Maedche and Staab [2], they consider noun phrases as concepts and in most cases they do not correspond to the names of the classes used in practice. Furthermore, it complicates the construction of an ontology in a situation in which ontology concepts are already available and also makes it harder to compare their results with those of other techniques.

PARNT is a more generalist solution to support specialists in finding non-taxonomic relationships in a wide range of pure text corpus and domains, when ontology concepts are available. It is possible because our proposal is parameterized and uses more adequate solutions for the extraction of *candidate relationships* and *refinement* phases (section V).

V. THE PARNT TECHNIQUE

PARNT is a semi-automatic LNTRO technique that uses NLP and statistical solutions to extract non-taxonomic relationships of predefined ontology concepts from an English corpus.

Because it is parameterized and provides new solutions like the apostrophe rule (AR) in the *extraction of relationships* phase and bag of labels in the *refinement* phase, PARNT is able to help developers getting good results in a wider range of situations when compared with most related works. The solutions adopted by PARNT for each phase of the generic process of LNTRO [9] illustrated in Fig. 1 are summarized in Table 1 and described in sections A to E.

TABLE I. PARNT SOLUTIONS FOR LNTRO

Phases		**PARNT solutions**
Corpus construction		Not approached
Extraction of candidate relationships	Corpus annotation	Tokenization, sentence splitter, morphological analysis, verb phrase chunking, and lemmatization
	Extraction of relationships	Sentence rule (SR), sentence rule with verb phrase (SRVP) and apostrophe rule (AR)
Refinement		Frequency of co-occurrence and Bag of labels
Evaluation by the specialist		Manual selection and edition of non-taxonomic relationships
Ontology update		Execution of the procedure to update the ontology file in owl format with non-taxonomic relationships.

A. Corpus construction

PARNT does not define a specific solution to be adopted in this phase and the specialist is the one responsible for choosing the one that best suits the needs for that situation. Manual [11] and automatic supported [14] [27] guidelines have been proposed for the construction of corpora. However, there is also the possibility of using a corpus already available in the case that the execution of this phase is dispensable.

B. Corpus annotation

This phase aims at adding annotations to the corpus. These annotations are needed for the application of the extraction rules selected by the expert in the *extraction of relationships* phase. PARNT applies five techniques of NLP (Fig. 2) executed in the order by which they are described in the following paragraphs.

Tokenization is a basic NLP task and its execution is a prerequisite for the application of any other NLP technique.

Sentence splitter is necessary because the sentence is the linguistic unit from which non-taxonomic relationships are extracted by applying the rules selected in the *extraction of relationships* phase. Lemmatization is used to improve the recall of the search for ontology concepts in the corpus. For example, the matching between the ontology concept *lawyer* and the term *lawyers* would not occur if the corpus was not lemmatized.

Morphological analysis classifies words in grammatical categories and is used in conjunction with verb phrase chunking to find verb phrases suggested as labels of the relationships. For example, the verb phrases *violates* and *can draw up* are labels for the relationship between the concepts *party* and *agreement* extracted from the following two sentences respectively: *If one party violates a settlement agreement the other may bring a lawsuit alleging a breach of contract* and *Although parties can draw up a separation agreement without the assistance of lawyers, it is often risky to do so*.These two NLP techniques are executed only if the SRVP rule (section C) is used in the *extraction of candidate relationships*.

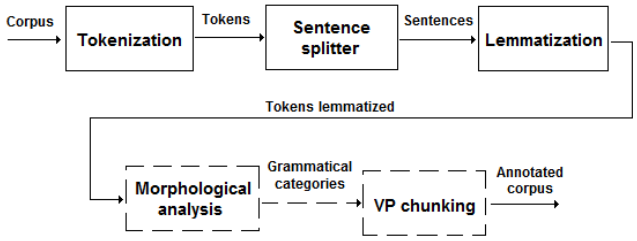

Figure 2. The Corpus annotation phase.

C. Extraction of candidate relationships

In this phase, a set of extraction rules selected by the specialist are used to extract candidate relationships from the previously annotated corpus. PARNT provides three types of extraction rules: the sentence rule (SR), the sentence rule with verb phrase (SR) and the apostrophe rule (AR). To illustrate their application, sentences from a corpus in the family law domain and the concepts "marriage", "spouse", "party", "child support" and "agreement" from its corresponding ontology will be used.

The SR extraction rule is based on the intuition that two consecutive concepts in the same sentence are probably non-taxonomically related. Considering the c_1 and c_2 ontology concepts, this rule can be formalized in the regular expression in PCRE (Perl Compatible Regular Expressions): $(?i)c_1((?!('s?)).)*?c_2$. Considering the sentence, *Parties can make agreements with respect to child support, which can be incorporated into a separation agreement,* the tuples <*party*, *agreement*>, <*agreement*, *child support*> and <*child support*, *agreement*> would be extracted.

The sentence rule with verb phrase (SRVP) considers that two consecutive concepts in the same sentence with a verb phrase (*vp*) between them are probably non-taxonomically related and can be formalized in the regular expression in PCRE: $(?i)c_1.*?vp.*?c_2$. This rule tends to provide lower recall than the SR one, because, for example, it cannot extract the tuple <*marriage, spouse*> from the sentence "The date and place of any previous marriages of either spouse as well as the date, place and circumstances under which they were terminated don't interfere in the present one.", which corresponds to a valid relationship. However, SRVP tends to offer higher precision.

The AR extraction rule is based on the intuition that two consecutive concepts with either strings "'s" or "'" between them have a high probability of being non-taxonomically related. The apostrophe rule can be formalized with the regular expression in PCRE: $(?i)c_1's?\ c_2$. For example, for the sentence "While the court will generally honor the parties' agreements as set forth in the separation agreement," the extracted tuple would be <*party, agreement*>.

The sets of tuples extracted by the extraction rules are represented in Fig. 3. The universe set (U) corresponds to all pairs of adjacent concepts present in each sentence in a corpus. The sub-sets T and F correspond to tuples that are true relationships and invalid ones respectively, whereas the sets SAR, SSRVP and SSR correspond to all tuples that are extracted with AR, SRVP and SR respectively.

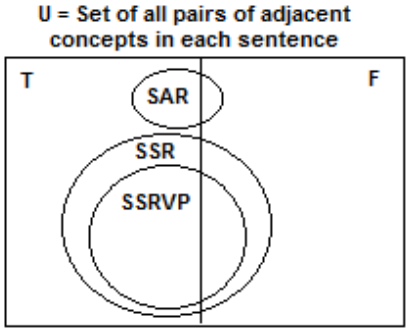

Figure 3. The sets of tuples obtained with the extraction rules.

Formally we can assert that SSR $\subseteq$ SSRVP, meaning that all extraction made by SRVP are also made by SR but not the reverse and that SAR $\cap$ SSR = $\varnothing$, which means that the relationships extracted by AR are not extracted by the SR, and vice versa. Furthermore, for most corpora, SAR < SSRVP < SSR. With respect to precision and recall the following can be stated: recall(SR) $\geq$ recall(SRVP) > recall(AR) and precision(SR) $\leq$ precision(SRVP) < precision(AR).

D. Refinement

PARNT provides two statistical solutions for this phase, the frequency of co-occurrence and bag of labels.

The general idea of bag of labels is to calculate the frequency of pairs of ontology concepts (<c_1, c_2>), independent of their order, and store the corresponding verb phrases in its bag of labels. The result is presented to the specialist that chooses the most appropriate verbal label for that relationship. This solution is used to filter the relationships extracted with SRVF extraction rule.

The frequency of co-occurrence calculates the frequency of pairs of ontology concepts (<c_1, c_2>), independent of their order. This solution is used to filter the relationships

extracted with AR or SR and therefore do not recommended labels to the suggested relationships. For both solutions the specialist can experimentally set the pruning parameter minimum frequency.

E. Evaluation by the Specialist and ontology update

No technique of NLP, ML or Statistic is capable of replacing the expert decision in an environment of ambiguous nature, as is the learning from natural language sources. Therefore, the goal of this phase is to make the best possible suggestions to the user and give him/her the control over the final decision. Thus, the result of the technique should be evaluated by a specialist before the relationships can be definitely added to the ontology. Issues such as the scope of the knowledge to be represented, the level of generalization, the real need of adding a relationship, its direction and label must ultimately be evaluated, selected, and possibly adjusted by an expert. Then a procedure to update the ontology owl file with these non-taxonomic relationships is executed.

VI. EVALUATING PARNT

To experimentally evaluate PARNT its recommendations were compared to those made by Villaverde et al. [12]. This technique was chosen because similarly to PARNT it presents the characteristics of having as input the ontology concepts and a corpus in English and also recommends labels to the relationships suggested to the specialist. Genia [7] corpus and ontology (version 3.0) were used as input. The corpus has 2000 documents, 18545 sentences and 436967 words, whereas the ontology has 47 concepts and 28 non-taxonomic relationships used as reference to calculate the evaluation measures recall and precision.

For this experiment PARNT was configured with SRVP and bag of labels for the *extraction of relationships* and *refinement*, respectively. The parameters minimum frequency, minimum support and minimum confidence of the refinement solutions of bag of labels and mining association rules were all set to zero. The recommendations were ordered by frequency of tuples for PARNT and by confidence for Villaverde et al. [12]. Recall (Fig. 2) and precision (Fig. 3) were calculated for the first sixty recommendations taken cumulatively in groups of ten.

Fig. 4 shows that PARNT had higher recall for the initial set of ten recommendations, and a lower growing rate with the increase in the number of recommendations. This desirable characteristic allows PARNT to perform a better separation between true and false relationships.

For the same initial set of ten recommendations Villaverde et al. [12] presented a lower value for recall and also a lower increasing rate than the one presented by PARNT for subsequent recommendations. This is due to the fact that true relationships are more uniformly distributed across the set of suggested relationships then those suggested by PARNT. This characteristic is undesirable since it makes the separation between true and false relationships more difficult. Furthermore, the specialist has to experimentally set two pruning parameters and not only one, as done with PARNT.

Fig. 5 shows that PARNT presented higher precision for the initial set of ten relationships than Villaverde et al. [12]. It was also noticeable a decreasing rate in precision with the increase in the number of recommendations. This confirms that PARNT makes a better separation between true and false non-taxonomic relationships.

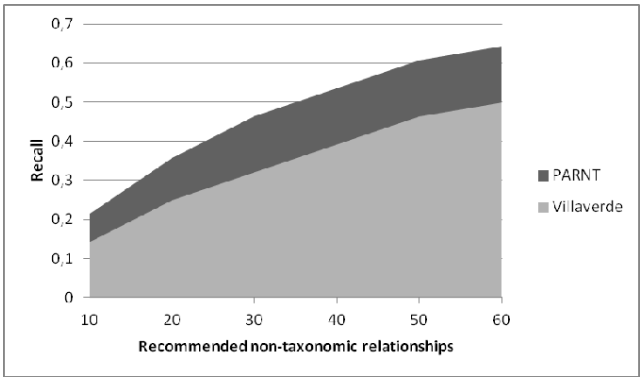

Figure 4. Recall values for PARNT and Villaverde et al.

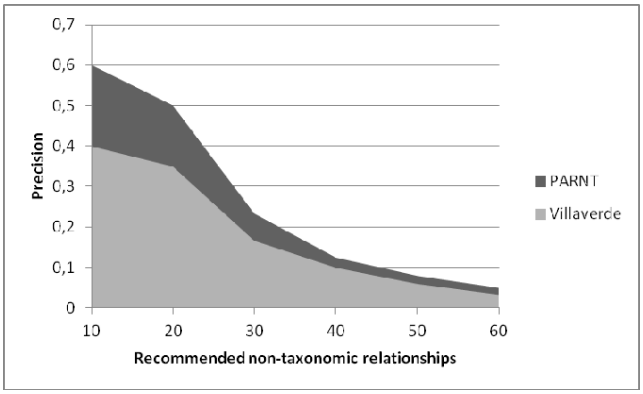

Figure 5. Precision values for PARNT and Villaverde et al.

VII. CONCLUDING REMARKS

LNTRO techniques as well as any other in the area of ontology learning are subject to a great amount of noise because the source from which information is extracted is unstructured. Thus highly customizable solutions are needed for these techniques to be applied to the widest possible range of situations.

This paper presented PARNT a semiautomatic technique for the extraction of non-taxonomic relationships of ontologies from pure text that uses NLP and statistic solutions. This technique presents several advantages, namely the control over the execution of its extraction rules, the AR and the bag of labels. The AR gives a differentiated treatment to extractions that have higher probability of being non-taxonomic relationships. As for the bag of labels, it is an origin solution to the refinement phase.

PARNT was experimentally evaluated comparing its effectiveness in terms of recall and precision to the proposal

of Villaverde et al. [12]. Initial results confirmed our hypotheses of PARNT and bag of labels being better solutions. However, more evaluations have to be conducted by comparing PARNT suggestions to those made by other LNTRO techniques for the same corpora and corresponding ontologies.

ACKNOWLEDGMENT

This work is supported by CNPq and CAPES, research funding agencies of the Brazilian government.

REFERENCES

[1] A. Fader, S. Soderland and O. Etzioni. Identifying Relations for Open Information Extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, 2011.

[2] A. Maedche and S. Staab. Mining non-taxonomic conceptual relations from text. In Knowledge Engineering and Knowledge Management. Methods, Models and Tools: 12th International Conference. Proceedings. pp. 189-202, 2000.

[3] C. Fellbaum. WordNet: An Electronic Lexical Database. Cambridge: MIT Press. pp. 23-24, 1998.

[4] D. Freitag. Information extraction from HTML: Application of a general machine learning approach. In Proceedings of the 15th Conference on Artificial Intelligence. pp. 517-523, 1998.

[5] D. Sanchez and A. Moreno. Learning non-taxonomic relationships from web documents for domain ontology construction. Data and Knowledge Engineering, 64(3), pp. 600-623, 2008.

[6] E. Sirin, J. Hendler, B. Parsia. Semi-automatic composition of web services using semantic descriptions. In: Proceedings of the ICEIS Workshop on Web Services: Modeling, Architecture and Infrastructure, 2002.

[7] F. Rinaldi, G. Schneider, K. Kaljurand, J. Dowdal, C. Andronis, A. Persidis and O. Konstanti. Mining relations in the GENIA corpus. Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics, pp. 61 - 68, 2004.

[8] I. Jurisica, J. Mylopoulos and E. Yu. Using ontologies for knowledge management: an information systems perspective. In: Knowledge: creation, organization and use – Proceedings of the 62nd annual meeting of the American Society for Information Science, Washington, DC, pp. 482–496, 1999.

[9] I. Serra, R. Girardi and P. Novais. The Problem of Learning Non-taxonomic Relationships of Ontologies from Text. In proceedings of the 9th Conference on Distributed Computing and Artificial Intelligence, Salamanca, Spain, pp. 485-492, 2012.

[10] J. Allen. Natural Language Understanding. Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc. 1995.

[11] J. Sinclair. Corpus creation. In Language, learning and community, eds. C Candlin and T McNamara, NCELTR Macquarie University, pp. 25-33, 1989.

[12] J. Villaverde, A. Persson, D. Godoy, A. Amandi. Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. Expert Syst. Appl. 36(7), pp. 10288-10294, 2009.

[13] K. Bontcheva and H. Cunningham. The Semantic Web: A New Opportunity and Challenge for Human Language Technology, In Proceedings of the Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference, Sanibel Island, 2003.

[14] M. Baroni and S. Bernardini. BootCaT: Bootstrapping corpora and terms from the web. In Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC). Lisbon: ELRA, pp. 1313-1316, 2004.

[15] N. Guarino, C. Masolo and C. Vetere. Ontoseek: Content-based Access to the web. IEEE Intelligent Systems, 14 (3), pp. 70-80, 1999.

[16] P. Buitelaar, P. Cimiano and P. Magnini. Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Amsterdam, The Netherlands, 2006.

[17] P. Cimiano, J. Volker and R. Studer. Ontologies on Demand? – A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text. In: Information, Wissenschaft und Praxis 57 (6-7), pp. 315-320, 2006.

[18] R. Dale, H. Moisl and H. L. Somers. Handbook of natural language processing. CRC, 2000.

[19] R. Girardi. Guiding Ontology Learning and Population by Knowledge System Goals. In: Proceedings of International Conference on Knowledge Engineering and Ontology Development, Ed. INSTIIC, Valence, pp. 480 – 484, 2010.

[20] R. Girardi and B. Ibrahim. Using English to retrieve software. Journal Of Systems Software: Special Issue on Software Reusability, New York, Elsevier. 30 (3), pp. 249 - 270, 1995.

[21] R. Girju, A. Badulescu and D. Moldovan. Learning semantic constraints for the automatic discovery of part-whole relations, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada, pp. 1-8, 2003.

[22] R. Srikant, R. Agrawal. Mining generalized association rules. In Proceedings of the International Confeference on Very Large Databases (VLDB 19 95), pp. 407-419, 1995.

[23] T. Finin, R. Fritzson, D. McKay, R. McEntire. KQML as an Agent Communication Language. In: Proceedings of the 3rd International Conference on Information and Knowledge management, pp. 456-463, 1994.

[24] T. Mitchell. Machine Learning, McGraw Hill, 1997.

[25] T. P. Mohamed, E. R. H. Junior and T. M. Mitchell. Discovering Relations between Noun Categories. In Proceedings of the conference on empirical methods in natural language processing (EMNLP 2011), Stroudsburg, Pennsylvania: Association for Computational Linguistics, pp. 1447-1455, 2011.

[26] T. R. Gruber. Toward Principles for the Design of Ontologies used for Knowledge Sharing, International Journal of Human-Computer Studies. Nº 43, pp. 907-928, 1995.

[27] W. H. Fletcher. Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton (eds.) Corpus Linguistics in North America. Amsterdam: Rodopi. pp.191-205, 2004.